Journal of Hydrology 564 (2018) 208–221

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

Development of non-parametric evolutionary algorithm for predicting soil moisture dynamics

Yongchul Shin[a], Binayak P. Mohanty[b,*], Amor V.M. Ines[b,1]

[a] School of Agricultural Civil & Bio-Industrial Engineering, Kyungpook National University, 80 Daehakro, Bukgu, Deagu 41566, Republic of Korea
[b] Department of Biological and Agricultural Engineering, Texas A&M University, 2117 TAMU, College Station, TX 77843, United States

ARTICLE INFO

This manuscript was handled by P. Kitanidis, Editor-in-Chief, with the assistance of Vahid Nourani, Associate Editor

Keywords:
HMMGA
Precipitation
Optimal state sequences (weather states)
Rainfall occurrence probabilities
Soil moisture dynamics

ABSTRACT

Prediction of soil moisture is critical for water resources management. With Global Precipitation Measurement (GPM) and Soil Moisture Active Passive (SMAP) satellites by NASA, spatio-temporal interrelation between rainfall and soil moisture fields (at different extents) will be of great value for satellite product calibration/validation and other hydrologic science investigations. In this study, we explored a non-parametric evolutionary algorithm for prediction of soil moisture from a time series of spatially-distributed rainfall across multiple weather locations under two different hydro-climatic regions. A new genetic algorithm-based hidden Markov model (HMMGA) was developed to estimate long-term soil moisture dynamics at different soil depths using precipitation data as a proxy. Also, we tested transposability of our approach across time under different climatic conditions. To test the new approach, we selected two different soil moisture fields, Oklahoma (130 km × 130 km) and Illinois (300 km × 500 km), during 1995–2009 and 1994–2010, respectively. We found that the newly developed framework performed well in predicting soil moisture dynamics at different spatial extents. Although our approach has limitations in predicting daily values, it estimates well the weekly soil moisture across the spatial and temporal domains with predictable uncertainties. Furthermore, this approach could provide advantages for good transposability under different weather conditions compared to those of physically-based hydrological models. Overall, our suggested approach could predict weekly soil moisture estimates with precipitation and soil moisture histories and showed the potential of transposability under different weather and land surface conditions. Since the proposed algorithm requires only precipitation (and historical soil moisture data) from existing, established weather stations, it can serve an attractive alternative that can forecast soil moisture using climate change scenarios.

1. Introduction

Soil moisture is an important component of many hydro-climatic processes. Accurate measurements of soil moisture are required for water resources management, understanding rainfall and runoff processes, irrigation scheduling, water quality monitoring, determining the partitioning of sensible and latent heat fluxes, etc. (Albergel et al., 2008; Barrett and Petropoulos, 2013; Brocca et al., 2011). In general, *in-situ* soil moisture measurements are assumed as fairly accurate observations, but these point-scale measurements may not be representative at field-scales. Since Remote Sensing (RS) soil moisture retrievals have been developed and improved, RS techniques are better suited for estimating near-surface soil moisture with respect to spatial and temporal coverage across the world (Ines and Mohanty, 2008a,b, 2009; Shin et al., 2013; Shin and Mohanty, 2013; Mohanty et al., 2017). The major concern regarding RS products is, however, the question of whether their resolution is too coarse to represent the scale at which hydrological processes occur (Engman, 1991; Entekhabi et al., 1999; Crow et al., 2012; Peng et al., 2017). This indicates that RS data suffer on account of the scale discrepancy between observation (RS) and model resolution. Furthermore, scaling down and up RS products requires the understanding of various complex stochastic and physical processes that can mimic a natural phenomenon such as the variability of land surface, the terrestrial ecosystem-atmospheric interactions, etc.

For obtaining soil moisture dynamics, various physically-based hydrological models such as Soil-Plant-Atmosphere-Water (SWAP) (van Dam et al., 1997), Community Land Model (CLM) (CLM3.0 User's Guide, Oleson et al., 2004), Variable Infiltration Capacity (VIC)

* Corresponding author.
E-mail address: bmohanty@tamu.edu (B.P. Mohanty).
[1] Department of Biosystems and Agricultural Engineering, Michigan State University, 1066 Bogue Street, 584B Plant and Soil Science, East Lansing, MI 48824, United States.

https://doi.org/10.1016/j.jhydrol.2018.07.003
Received 26 March 2018; Received in revised form 28 June 2018; Accepted 2 July 2018
Available online 04 July 2018
0022-1694/ © 2018 Elsevier B.V. All rights reserved.

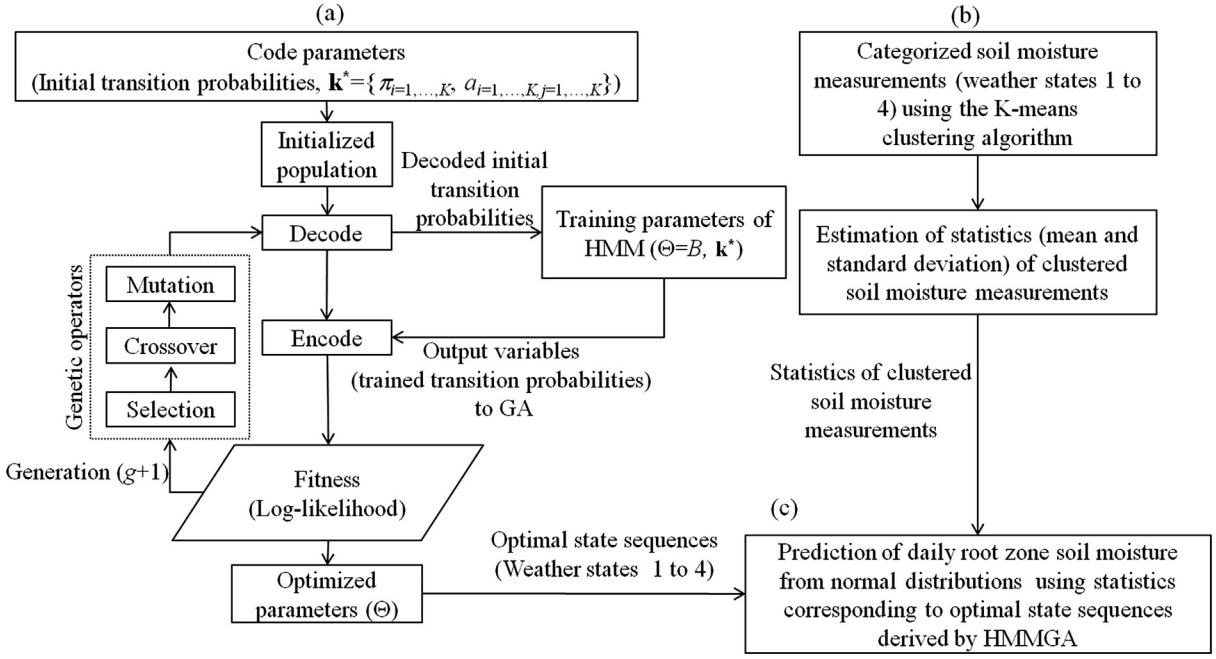
(a)
Code parameters
(Initial transition probabilities, $\mathbf{k}^*=\{\pi_{i=1,\ldots,K}, a_{i=1,\ldots,K,j=1,\ldots,K}\}$)
Initialized population
Decoded initial transition probabilities
Training parameters of HMM ($\Theta=B$, $\mathbf{k}^*$)
Decode
Encode
Genetic operators
Mutation
Crossover
Selection
Output variables (trained transition probabilities) to GA
Generation (g+1)
Fitness (Log-likelihood)
Optimized parameters (Θ)
(b)
Categorized soil moisture measurements (weather states 1 to 4) using the K-means clustering algorithm
Estimation of statistics (mean and standard deviation) of clustered soil moisture measurements
Statistics of clustered soil moisture measurements
(c)
Optimal state sequences (Weather states 1 to 4)
Prediction of daily root zone soil moisture from normal distributions using statistics corresponding to optimal state sequences derived by HMMGA

Fig. 1. Flow chart of the multivariate framework for forecasting soil moisture dynamics; (a) hidden Markov model based on a genetic algorithm (HMMGA), (b) categorized soil moisture measurements using the K-means clustering algorithm, and (c) predictions of soil moisture dynamics. $\pi_{i,j}$: the initial hidden (weather) states. $a_{i,j}$: the transition matrix of hidden (weather) states. B: the output probability (rain or no rain) matrix ($b_j(o)$). Θ: the trained parameter set of B and $\mathbf{k}^*$ ($\pi_{i,j}$, $a_{i,j}$). g: the number of given generations for genetic algorithm. State 1: wet condition; State 2: relatively wet condition; State 3: relatively dry condition; State 4: dry condition.

Macroscale Hydrologic Model (Liang et al., 1996), the U.S. Department of Agriculture Hydrograph Laboratory (USDAHL) model (Holtan, 1975), the Sacramento Soil Moisture Accounting (SAC-SMA) model (Peck, 1976), among others were also developed. Besides the above mentioned deterministic models, several other approaches have been developed for soil moisture estimations. A sequential assimilation approach was developed (Mahfouf, 1990) and improved (Bouttier et al., 1993a,b) for soil moisture estimations from atmospheric temperature and relative humidity. A genetic algorithm-based assimilation scheme was developed for quantifying effective soil hydraulic properties based on inverse method (IM) (Ines and Mohanty, 2008a,b,2009; Shin et al., 2012, 2013; Shin and Mohanty, 2013). However, these approaches could be limited by the number and complexity of required physical parameters in modeling. Especially, various environmental conditions (i.e., initial and bottom boundary conditions, soil textures, vegetation, atmospheric forcings, topography, etc.) influence the performances of hydrological models. This may indicate that predictive simulation outputs are dependent on model structures and user's modeling skills. Furthermore, these hydrological models also have drawbacks in capturing model parameters that are suitable for both current (calibration) and projection (validation) conditions (Nazemi et al., 2006; Seiller et al., 2012; Dumedah et al., 2012). Common hydrological models usually assume that parameters estimated during calibration periods (associated to hydro-climatic conditions) are applicable to validation periods under the assumption of the stationarity of rainfall-runoff transformation. This assumption is only valid when environmental (weather) conditions are not much different from calibration periods, but in a climate change context, contrasted weather conditions between calibration and validation periods may bring out the question of the stationarity hypothesis. Additionally, a simple analytical approach (Pan et al., 2003) based on a linear stochastic partial differential equation (Entekhabi and Rodriguez-Iturbe, 1994) has been developed for estimating surface soil moisture using precipitation data. However, this approach also requires land surface information for determining loss coefficient, infiltration coefficient, and averaging window size and provides only linear equation-based surface soil moisture estimates during relatively short-term periods. Few stochastic approaches (e.g., Hidden Markov Model-HMM (Kirshner, 2005), Nonstationary Hidden Markov Model-NHMM (Robertson et al., 2004; Mehrotra and Sharma, 2005; Kwon et al., 2017), etc.) have been used to downscale satellite-based products such as Advanced Microwave Scanning Radiometer 2-AMSR2 (JAXA, 2013)/Soil Moisture Ocean Salinity-SMOS (Kerr et al., 2012)/Soil Moisture Active Passive-SMAP (Das et al., 2011) soil moisture, Tropical Rainfall Measuring Mission-TRMM (Kummerow et al., 1998), Global Precipitation Measurement-GPM (Hou et al., 2014), etc. To date, no stochastic methods have been designed for forecasting long-term soil moisture dynamics at the spatio-temporal domains across the land surface using other proxies.

In this study, we explored a non-parametric evolutionary algorithm for forecasting long-term weekly soil moisture dynamics at the near-surface and deeper soil layers using a stochastic optimization approach on a multivariate time series of rainfall datasets obtained from a network of weather stations. The objective of our study was to predict spatially and temporally distributed weekly soil moisture dynamics based on rainfall history and weather dry/wet patterns using a newly developed genetic-algorithm based hidden Markov model (HMMGA) and assess its transposability across time under different hydro-climatic regions.

2. Materials and methods

2.1. Conceptual multivariate framework

In this study, we developed a novel methodology for weekly soil moisture predictions at different study areas using multi-site precipitation data. The approach comprised of three steps; (i) development of a hidden Markov model using a genetic algorithm (HMMGA) for estimating the optimal hidden state sequence (representing weekly weather states) using precipitation; (ii) estimation of statistics (mean and standard deviation) of categorized soil moisture measurements based on the weather states, and iii) soil moisture predictions using the derived hidden (weather) states and statistics of categorized soil moisture history. Note that the weather states indicate the hidden states (State 1: wet, State 2: relatively wet, State 3: relatively dry, and State 4:

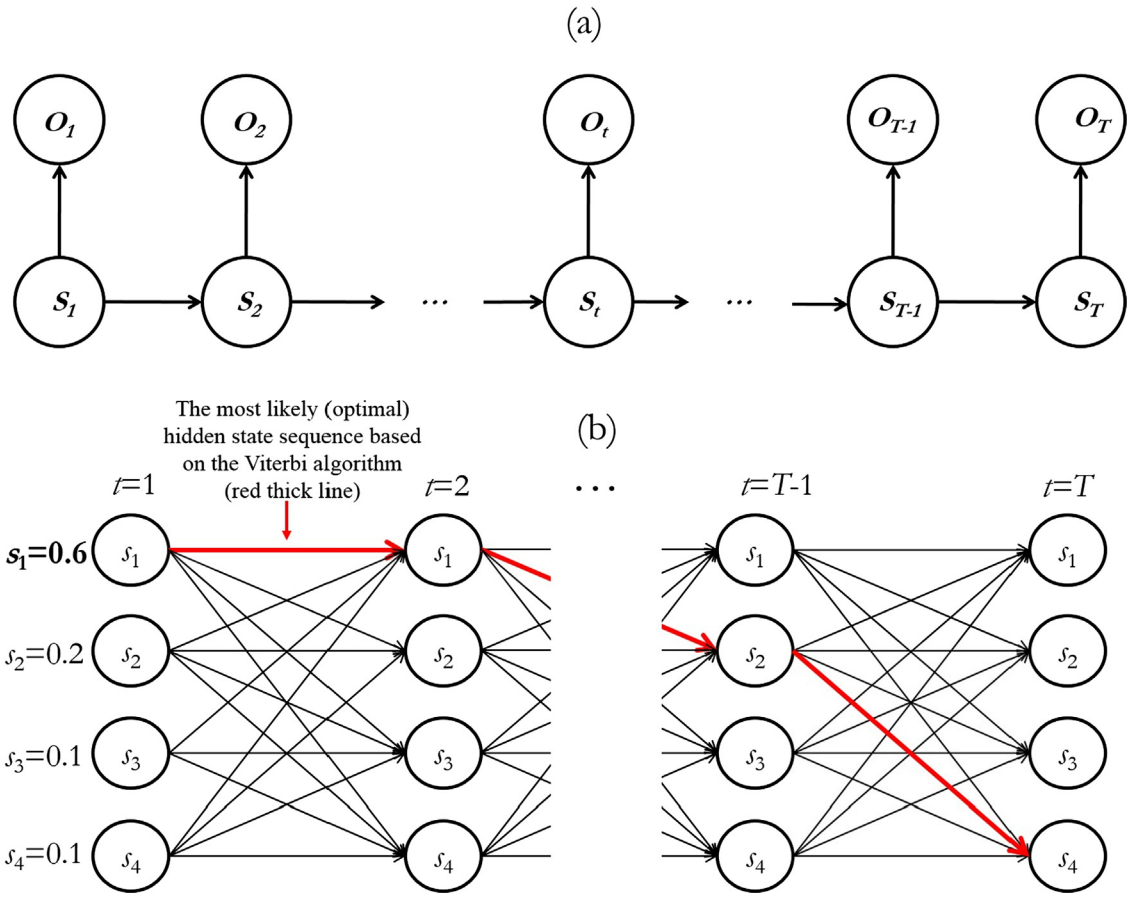

Fig. 2. (a) Graphical model representation of a hidden Markov model; (b) the optimal hidden (weather) state paths (red thick lines) using the Viterbi algorithm. O_t: the probabilities (outputs) of rainfall occurrences with the time series (t). S_t: the hidden (weather) state vector comprised of different weather states ($s_{1:4}$) with the time series (t). $s_{1:4}$: the four hidden (weather) states (State 1: the wet condition, State 2: relatively wet condition, State: 3 relatively dry condition, and State 4: dry condition) with the time series (t).

dry conditions) derived by HMMGA. The methodology is schematically shown in Fig. 1.

Stochastic approaches coupled with physically-based hydrological models have been commonly used in hydrological predictions such as optimum interpolation-based sequential assimilation technique for estimating soil moisture (Bouttier et al., 1993a,b), genetic algorithm-based near-surface (Ines and Mohanty, 2008a,b, 2009) and layer-specific (Shin et al., 2012) soil moisture data assimilation, among others. However, these coupled models are parameter intensive and are considerably influenced by the initial parameter values. Also, these hydrological models have drawbacks in meeting the transposability for calibration and validation under contrasted weather conditions. In order to address these limitations of existing schemes, we developed a new algorithm by integrating a stochastic hidden Markov model (HMM) with genetic algorithm. The hypothesis of our proposed approach is that a HMM can provide better trained parameters for obtaining the hidden (weather) states through optimized initial parameters derived by a GA. This coupling eliminates the need of using a parameter intensive physical model and minimizes not only user's skills in determining initial parameter values, but also model structural (i.e., parameterizations, initial/bottom boundary conditions, governing equation, transposability, etc.) and measurement (i.e., land covers, soil information, etc.) errors.

2.1.1. Hidden Markov model (HMM) based on genetic algorithm

In a regular Markov model, state transition probabilities are directly visible to the observer. However, as its name suggests, a hidden Markov model (HMM) has invisible (or hidden) state transition probabilities. The notation of O_t is the M (multiple weather stations)-dimensional vector of measurements at time t, and the notation of $O_{1:T}$ ($O_1,\ldots,O_T$) is the sequential vectors of time length (T). The notation of $S_{1:T}$ ($S_1,\ldots,S_T$) is the corresponding sequential vectors of hidden (weather) states with each of the hidden states (S_t, $t = 1,\ldots,T$) taking on one of the hidden states (1 to K, note that K is the number of hidden states). Here, the hidden state vector of S_t at time t is comprised of different weather states ($S_t = \{s_{1,\ldots,K}\}$). Also, the observed rainfall value of $o_t^m = 1$ indicates "rain" on day t while $o_t^m = 0$ means "no rain" on day t at the weather station ($m = 1,\ldots,M$), respectively.

A HMM can be defined as a joint distribution on a rainfall occurrence observation (output) and hidden state of rainfall along the time (t) based on two conditional independence assumptions (Hughes and Guttorp, 1999; Robertson et al., 2004; Kirshner, 2005).

(i) In the first assumption, the multivariate rainfall observations (O_t) at time t are independent of all other variables across time t and conditional on the hidden (weather) state S_t at time t in Eq. (1),

$$P(O_t|S_{1:t}, O_{1:t-1}) = P(O_t|S_t) \quad (1)$$

(ii) In the second assumption, the hidden (weather) state process ($S_{1:T}$) is the first-order Markovian (representing that Markov process is homogeneous with the transition matrix of $K \times K$ in time) in Eq. (2).

$$P(S_t|S_{1:t-1}) = P(S_t|S_{t-1}) \quad (2)$$

In order to estimate $P(O_t|S_t)$, we simplified the assumption that the rainfall observations across the weather stations (m) at time t are

spatially independent and conditional on the hidden state (Robertson et al., 2004; Kirshner, 2005) as,

$$P(O_t{=}o|S_t = s) = \prod_{m=1}^{M} P(O_t^m{=}o|S_t = s\,){=} \prod_{m=1}^{M} p_{s,o}^m \tag{3}$$

$$\left.\begin{aligned} a_{ij} &= P(S_{t+1} = s_j|S_t = s_i,\ 1 \leqslant i, j \leqslant K) \\ &\sum_{j=1}^{K} a_{ij} = 1 \end{aligned}\right\} \tag{4}$$

$$\left.\begin{aligned} b_j(o) &= P(o_t = o|S_t = s_j),\ \ 1 \leqslant j \leqslant K,\ \ 0 \leqslant o \leqslant 1 \\ &\sum_{o=0}^{1} b_j(o) = 1 \end{aligned}\right\} \tag{5}$$

where the rainfall occurrences at each weather station (m) were denoted as "o" ($p_{s,o}^m$ means no rain for $o = 0$ and rain for $o = 1$). a_{ij} is the transition probability matrix on row state i and column state j, and $b_j(o)$ is the output probability matrix (note that $b_j(o)$ is the probability density function that "o" can be observed at state s_j). The readers are referred to Rabiner (1989), MacDonald and Zucchini (1997), Robertson et al. (2004), and Kirshner (2005) for a more detailed explanation of the HMM.

Sequences of daily rainfall occurrences (O_t) and hidden rainfall states (S_t) are shown in Fig. 2a. In this study, a GA searches the optimized first (initial) hidden state parameters ($\pi_{i\,=\,1,\cdots,K}$) at the time ($t = 1$) on row state i and transition matrix ($a_{i=1,\cdots K,j=1,\cdots,K}$) composed of parameters ($\mathbf{k}^* = \{\pi_i, a_{i,j}\}$) to be trained by a HMM, because GAs are powerful search algorithms for solving complex problems (parameters) through the survival of the fittest mechanism (Holland, 1975; Goldberg, 1989). Thus, a HMM can train a set of parameters (Θ, Eq. (6) for estimating the weather state conditions using the optimized initial hidden states ($\mathbf{k}^*$) provided by a GA,

$$\Theta = (B, \mathbf{k}^*) = (A, B, \pi_i) \tag{6}$$

where π_t is the hidden state parameters at time t, A is the transition probability matrix (a_{ij}) on row state i and column state j, and $B_{o=0,1\times T}$ is the output probability matrix ($b_j(o)$), respectively.

The appropriate number of transition probabilities ($K = 2,\ldots,N$) are necessary to predict accurate rainfall occurrences. It must be considered that a high N value will increase the number of elements in the transition matrix (which is of size $K \times K$) causing excessive repetitions of simulations. There are no set guidelines for determining the optimum orders of K. Robertson et al. (2004) tested the HMM with K ranging from 2 to 6 and suggested that four hidden states ($K = 4$) provide the most reasonable results. Furthermore, increasing hidden states require more soil moisture measurements to obtain statistics of averages and standard deviations corresponding to each hidden states. Therefore, we tested different orders of K (2,…,6) for selecting the optimum order K of hidden states, which gives better results.

Given the hidden states K, the parameters Θ of the HMM can be trained by searching for parameters that fit to the observed data. For this purpose, we employ the well-known maximum likelihood (l) approach based on the Baum-Welch algorithm (i.e., Rabiner (1989)) as shown in Eq. (7). This approach maximizes the conditional probability of the parameters based on the observed data.

$$l(\Theta) = \log P(O|\Theta) = \log \sum_{S_{1:T}} [P(S_1)\prod_{t=2}^{T} P(S_t|S_{t-1})]\ [\prod_{t=1}^{T} P(O_t|S_t)] \tag{7}$$

The Baum-Welch algorithm maximizes the log-likelihood of $l(\Theta)$ as a variant of an iterative Expectation-Maximization (EM, Dempster et al., 1977; Robertson et al., 2004) algorithm to obtain the maximum log-likelihood parameter set for models with hidden state variables. More details are referred to the Appendix A (Robertson et al., 2004). To compute the log-likelihood of the data $P(O_{1:T}|S_{1:T})$, the *Forward-Backward* procedure (Rabiner, 1989) was adopted as a recursive method. The *Forward* procedure means that the probabilities of hidden state (S_t) are calculated from the time from 1 to T, and then the probabilities of hidden states (S_t) are calculated from the time of T to 1 as the *Backward* procedure. The maximization process is finished when the probabilities derived by Forward and Backward procedures are matched. Then, the hidden state sequences can be obtained across the time ($t = 1,\ldots,T$). However, the possible probabilities of obtained hidden states will be exponentially increased along the time step ($t = 1,\ldots,T$) because each hidden value (s_1,s_2, …, or, s_K) of $S_t = \{s_1,\ldots,s_K\}$ at the time (t) has the possible probabilities (up to K), which will be transitioned to the next time ($t + 1$). For this reason, we need to find the most likely (optimal) hidden state sequence representing the daily weather state for the observation at every time steps ($t = 1,\ldots T$). Thus, our approach was integrated with the Viterbi algorithm (Viterbi path, Viterbi (1967)) to find the most likely hidden state sequence at every time steps in Eq. (9). The Viterbi algorithm find the highest probability among the hidden states (s_1,s_2, …, or, s_K) at the time (t), and then the possible hidden state probabilities are calculated at the next time step ($t + 1$). Thus, the highest (most likely) hidden state sequence can be calculated across the time (1 to T) as the *Forward* procedure, and then the Viterbi algorithm calculates the highest hidden state sequence across the time (T to 1) as the *Backward* procedure. When the most likely hidden state sequences obtained in the *Forward-Backward* procedures are matched, the most likely (optimal) hidden state sequence ($S_1,\ldots S_T$) that can represent rainfall occurrences ($O_1,\ldots,O_T$) can be determined through the joint distribution as,

$$V_1(i) = P(O_1|S_1 = \{s_i\})\pi_i \tag{8}$$

$$V_t(j) = \max_{1\leqslant i\leqslant K} P(O_t|S_t = \{s_i\})a_{ij}V_{t-1}(i) \tag{9}$$

Here V_t is the most probable state sequence probabilities and S_t is the most likely (optimal) state sequence representing the daily weather state for the observation at time t as shown in Fig. 2b. In the *Backward* procedure, the optimal state sequence can be retrieved by saving back pointers that remember which state S_t was used,

$$S_T = \max_{1\leqslant i\leqslant K} V_T(i) \tag{10}$$

$$S_t = S_{t+1}\mathrm{argmax}_{1\leqslant i\leqslant K} a_{ij}V_t(i) \tag{11}$$

Note that "arg max" meant the argument for selecting the maximum value in the given data sets ($i = 2,\ldots,K$). We tested the log-likelihood values (l: −468.0, −435.9, −429.0, −419.0, and −419.0) and correlations (R: 0.350, 0.380, 0.945, 0.873, and 0.873, see Section 2.3) between the observed rainfall occurrence probabilities and weather states derived by the HMMGA for various orders of hidden states ($K = 2,\ldots,6$) in the Oklahoma domain (see Section 2.2). Note that a log-likelihood value closer to zero denotes better performance. The log-likelihood values considerably decreased from $K = 2$ to $K = 5$, but the decreasing trend from $K = 5$ to $K = 6$ was relatively gentle. The correlation was highest at $K = 4$ indicating that four hidden states can represent the weather states better than others as shown in the findings of Robertson et al. (2004). Thus, we selected the four hidden states ($K = 4$) for estimating weather states. In this study, it is assumed that four hidden states represent the weather wet (state 1), relatively wet (state 2), relatively dry (state 3), and dry (state 4) conditions, respectively.

In this study, a modified-*micro*GA (Ines and Honda, 2005; Ines and Mohanty, 2008a) was used to search parameters ($\mathbf{k}^*$) along the generations. The parameters are composed of genes, which are arranged in an array of binary values called a chromosome. The modified-*micro*GA is a GA variant that uses a micro-population for searching parameters (Ines and Droogers, 2002; Carroll et al., 1996; Goldberg, 2002; Krishnakumar, 1989). It has the unique ability to restart when the searched chromosomes of the micro-population are similar in structure (90%) to each other before the generations are completed. This improves the algorithm computationally and makes it more efficient.

Table 1
The search spaces of variables and probabilities of selection, crossover, and mutation used in the genetic algorithm.

Locations	Parameters	Search space of parameters		No. of bit (L)	2L	Population size	Seed number	Number of generation	P_{creep}	P_{mutate}	P_{cross}
		Minimum values	Maximum Values								
Oklahoma domain	π_i	0.0	1.0	7	128	10	3000	30	0.3	0.3	0.3
	a_{ij}	0.0	1.0	7	128						
Illinois domain	π_i	0.0	1.0	7	128	10	3000	30	0.05	0.05	0.05
	a_{ij}	0.0	1.0	7	128						

$\pi_{i=1,\ldots,K}$: initial state probabilities of being in state i (K is the number of transition probabilities).
a_{ij}: transition probabilities ($K \times K$) of transitioning from state i to state j.
P_{creep}: creep probability.
P_{mutate}: mutation probability.
P_{cross}: crossover probability.

Restarting allows the modified-*micro*GA to search the solutions in the global space. In order to infuse new genes, creep (at the decimal level, D. L. Carroll, Fortran genetic algorithm (GA) driver, available at www.cuaerospace.com/carroll/ga.html) and intermittent jump (Ines and Mohanty, 2008a) mutations are implemented in the modified-*micro*GA. Furthermore, a time saving mechanism used in the GA remembers the elite (fittest) chromosome in the previous generation ($g-1$) (Ines and Honda, 2005). Then the elite chromosome is reproduced in the next generation (g) without computation (note that the modified-*micro*GA always reproduces the elite chromosome in the next generation). The search spaces of variables and probabilities of selection, crossover, and mutation used in a GA for the different study domains are shown in Table 1.

2.1.2. Categorizing soil moisture

In this study, HMMGA predicted the weather states using only the observed rainfall occurrences. However, soil moisture values are usually affected by the precipitation, as well as by the land surface conditions (i.e., soil textures, vegetation covers, topography, etc.) in different regions. Thus, corresponding to the four-hidden states fitted by HMMGA in Section 2.1.1, we classified historical soil moisture data into four categories ($K = 4$) ranging from wet to dry (represented by 1, 2, 3, 4 in order of weather states) at each weather station using the K-means clustering algorithm (Hartigan and Wong, 1979) based on the Euclidean distance,

$$d(x,y) = \frac{1}{W}\sum_{w=1}^{W}(x_w - y_w)^2 \tag{12}$$

where $d(x,y)$ is the Euclidean distance between two points of x_w and y_w on the order of w, x_w and y_w are the soil moisture measurement on the order of w, w is the running index, and W is the total number of soil moisture measurements, respectively.

Initially, the K-means clustering algorithm assigns the soil moisture data randomly to a cluster. This clustering approach repeats a two-step process in the given number ($n = 100$) of runs as follows; (i) the mean soil moisture data of individual cluster is estimated and (ii) the soil moisture data are reallocated to each cluster whose mean value is closest to the soil moisture data. For more details on the popular K-means clustering algorithm refer to Jain et al. (1999). In order to incorporate site specific soil moisture dynamics, the clustering of soil moisture data was done for each weather station (m) separately. Then, we calculated mean ($\mu_{i=1,\ldots,K,m=1,\ldots,M}$) and standard deviation (SD, $\sigma_{i=1,\ldots,K,m=1,\ldots,M}$) of the classified soil moisture measurements for each weather station. The soil wetness ranges (maximum and minimum) for four weather state categories were determined based on clustered soil moisture measurements at individual weather station.

2.1.3. Prediction of soil moisture

Based on the derived optimal state sequences ($S_1,\ldots,S_T$, Section 2.1.1) and estimated statistics ($\mu_{i,m}$ and $\sigma_{i,m}$, the Section 2.1.2) of historical soil moisture data, we generated normal distributions ($N(\mu_{i,m}, \sigma^2_{i,m})$) of soil moisture for each individual weather states along the time series (Casella and Berger, 2001; Kwon et al., 2017). The predicted daily soil moisture (θ_t^m) value was determined by randomly selecting a value from the generated normal distribution corresponding to the optimal state sequences selected by HMMGA (Eq. (13)). Note that the daily soil moisture predictions were averaged for deriving the weekly soil moisture values. Each optimal state value indicates the weather state (S_t) at time (t). Note that the function of $f()$ returns the randomly generated soil moisture values for each weather station (m) along the time series (t).

$$\theta_t^m = f(S_t,\ N(\mu_{i,m}, \sigma^2_{i,m}))\ \ \forall\ m\ \ \forall\ t \tag{13}$$

2.2. Study sites and data description

Since the proposed approach is sensitive to the domain size, two domains of different extents were chosen to evaluate the applicability of this approach under two hydro-climatic regions; (i) the Oklahoma domain (130 km × 130 km) with humid subtropical climate and (ii) the Illinois domain (300 km × 500 km) with humid continental climate, respectively. For the Oklahoma domain, we selected the seven Mesonet weather stations (http://www.mesonet.org/index.php) such as site 1-ELRE (silt loam), site 2-SPEN (sandy loam), site 3-MINC (silt loam), site 4-ACME (sandy loam), site 5-WASH (loam), site 6-PAUL (silt loam), and site 7-BYAR (sandy loam) in Fig. 3a. The weather stations have elevations in the range of 265.8–421.2 m above MSL. Sites 1, 3, and 4 in the west have higher elevations (404.3–421.2 m) than those located in the east (265.6–349.5 m). For the Illinois domain, we selected seven weather station sites from the International Soil Moisture Network (ISMN: http://www.ipf.tuwien.ac.at/insitu/) and Illinois State Water Survey (ISWS: https://www.isws.illinois.edu/) in Illinois such as site 1-Bondville (silt loam/silty clay loam), site 2-Dekalb (silty loam/silty clay loam), site 3-Freeport (silt loam), site 4-Monmouth (silt loam), site 5-Olney (silt loam), site 6-Orr Center (silt loam), and site 7-Peoria (silt loam) in Fig. 3b. The weather stations located in the northern portion of this domain had slightly higher elevations. The Oklahoma and Illinois areas generally have a flat topography. The soils and vegetation vary across the Oklahoma and Illinois regions, but the soil moisture observation were usually measured with native grass covers.

Daily precipitation datasets were obtained from the Oklahoma mesonet stations (1995–2009) in Oklahoma and the ISMN (1994–2010) in Illinois during the crop growing season (90 days from March 1 to May 29). At the Oklahoma sites, daily soil temperature readings (available from 2003 to 2009) measured by Campbell scientific 229-L sensors were converted to (0–5 cm) volumetric soil moisture with the empirical coefficients provided from the mesonet stations (see details, Illston et al., 2008). At the Illinois locations, biweekly volumetric soil moisture data at different soil depths (0–10, 10–30, 30–50, and 0–50 cm) were measured from 1994 to 2004 using Troxler neutron

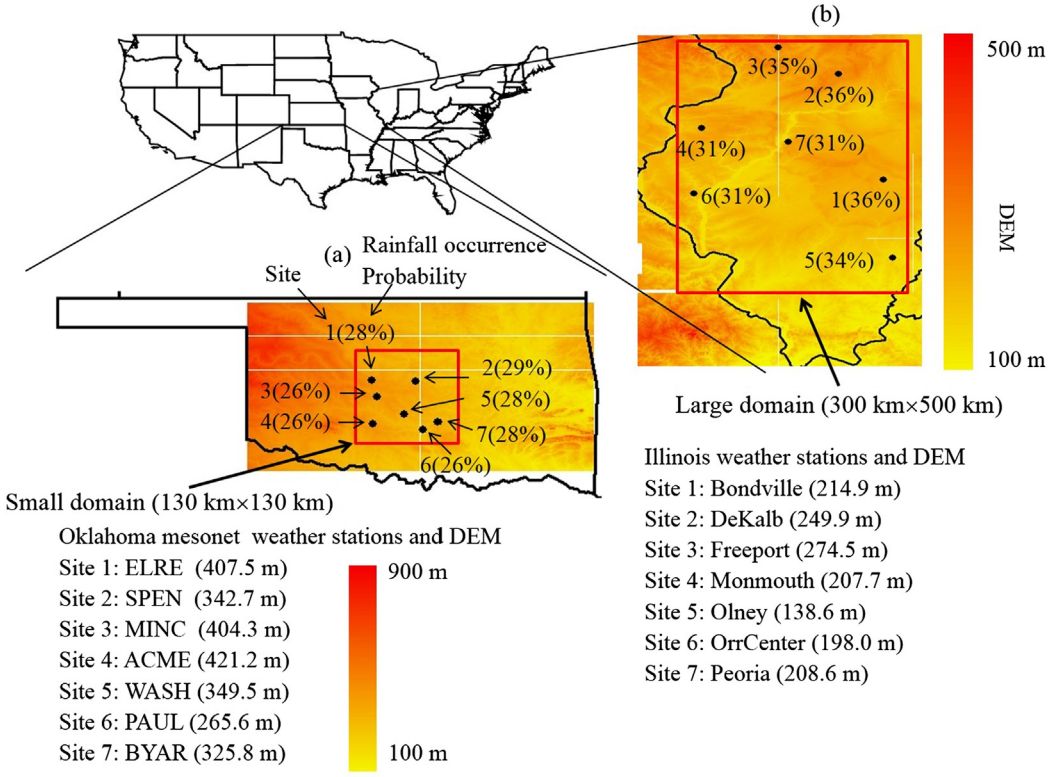

Fig. 3. Network of multiple locations (sites 1–7), digital elevation model (DEM), and observed rainfall occurrence probabilities under different hydro-climatic regions; (a) Oklahoma domain and (b) Illinois domain.

Table 2
Statistics of categorized soil moisture measurements at the soil depth (0–5 cm) using the K-means clustering algorithm in Oklahoma (2003–2009).

Soil depth	Sites	Oklahoma[a]															
		State 1 (Wet condition)				State 2 (Relatively wet condition)				State 3 (Relatively dry condition)				State 4 (Dry condition)			
		Mean	SD	Min.	Max.	Mean	SD	Min.	Max.	Mean	SD	Min.	Max.	Mean	SD	Min.	Max.
0–5 cm	1	0.391	0.014	0.379	0.418	0.350	0.013	0.328	0.361	0.293	0.015	0.273	0.314	0.248	0.011	0.234	0.266
	2	0.221	0.001	0.221	0.227	0.211	0.002	0.208	0.216	0.201	0.003	0.195	0.205	0.187	0.003	0.183	0.193
	3	0.364	0.014	0.346	0.391	0.321	0.007	0.316	0.331	0.290	0.012	0.260	0.302	0.224	0.011	0.215	0.253
	4	0.246	0.004	0.243	0.257	0.235	0.002	0.232	0.237	0.221	0.005	0.213	0.227	0.200	0.005	0.194	0.210
	5	0.294	0.003	0.293	0.304	0.281	0.003	0.274	0.283	0.253	0.009	0.238	0.265	0.217	0.011	0.201	0.233
	6	0.401	0.005	0.400	0.423	0.370	0.010	0.355	0.377	0.314	0.015	0.297	0.334	0.253	0.026	0.209	0.282
	7	0.204	0.000	0.202	0.206	0.198	0.002	0.195	0.199	0.186	0.003	0.181	0.191	0.173	0.004	0.169	0.179

[a] Campbell scientific 229-L sensor (daily soil temperature estimates were measured by this sensor and converted to daily soil moisture dynamics using the equation of volumetric water content (VWC: cm^3 cm^{-3}) with the empirical coefficients provided from the mesonet stations (Illston et al., 2008)).

surface and depth probes by the Illinois State Water Survey (Hollinger and Isard, 1994). Note that soil moisture at 0–50 cm was obtained by averaging soil moisture measurements at the soil depths of 0–10, 10–30 and 30–50 cm. Descriptive statistics (mean, standard deviation, minimum, and maximum) of soil moisture measurements for the Oklahoma and Illinois sites were shown in Table 2 and 3. Precipitation data from the weather stations were available from 1995 to 2009 in Oklahoma and 1994–2010 in Illinois. However, the soil moisture data were measured for a shorter duration (2003–2009 for Oklahoma and 1994–2004 for Illinois). The optimal (hidden) state sequences for both study domains were determined by the HMMGA using the entire range of precipitation data available. However, in order to generate the normal distributions for the soil moisture, we used only the estimated optimal state sequences from 2003 to 2009 for Oklahoma and 1994–2004 for Illinois. The measurements used for the calibration were taken from 1994 to 2000 and the validation was conducted from 2001 to 2004 at the Illinois sites.

2.3. Analysis method

The challenge of this study is to model and predict the rainfall occurrence probabilities and soil moisture dynamics by using historical precipitation data and statistics of soil moisture measurements across the multiple locations (sites 1–7) for the Oklahoma and Illinois domains. In this study, the spatial dependence of rainfall occurrence was analyzed during given time periods in different study domains.

To verify the accuracy of our suggested HMMGA method, we compared soil moisture predictions derived by this approach with those from a near-surface soil moisture assimilation scheme based on inversion model (IM) (Ines and Mohanty, 2008a) in Illinois. The data assimilation scheme determined effective soil hydraulic properties (α, n, θ_{res}, θ_{sab} K_{sat}) by tuning a physical-based hydrological (Soil Water Atmosphere Plant: SWAP) model with a genetic algorithm (see details in Ines and Mohanty, 2008a). Here, the soil hydraulic variables of α (shape parameter equivalent to the inverse of the bubbling pressure), n

Table 3
Statistics of categorized soil moisture measurements at the different soil depths (0–10, 10–30, 30–50, and 0–50 cm) using the K-means clustering algorithm in Illinois (1994–2000).

Soil depths	Sites	Illinois[a]															
		State 1 (Wet condition)				State 2 (Relatively wet condition)				State 3 (Relatively dry condition)				State 4 (Dry condition)			
		Mean	SD	Min.	Max.	Mean	SD	Min.	Max.	Mean	SD	Min.	Max.	Mean	SD	Min.	Max.
0–10 cm	1	0.493	0.037	0.456	0.558	0.414	0.018	0.384	0.439	0.338	0.020	0.312	0.363	0.273	0.024	0.218	0.300
	2	0.444	0.025	0.415	0.503	0.380	0.016	0.354	0.405	0.316	0.016	0.298	0.343	0.260	0.015	0.244	0.280
	3	0.448	0.016	0.434	0.481	0.405	0.010	0.389	0.421	0.339	0.016	0.304	0.361	0.246	0.015	0.221	0.265
	4	0.464	0.032	0.424	0.506	0.380	0.020	0.352	0.417	0.311	0.021	0.281	0.335	0.240	0.029	0.191	0.262
	5	0.531	0.025	0.509	0.581	0.474	0.017	0.450	0.494	0.420	0.014	0.399	0.445	0.352	0.036	0.306	0.384
	6	0.448	0.023	0.417	0.482	0.374	0.017	0.348	0.406	0.310	0.021	0.271	0.339	0.202	0.037	0.167	0.240
	7	0.515	0.044	0.473	0.580	0.404	0.020	0.373	0.441	0.335	0.018	0.304	0.361	0.252	0.030	0.206	0.287
10–30 cm	1	0.428	0.012	0.416	0.447	0.385	0.011	0.369	0.400	0.349	0.011	0.331	0.365	0.285	0.015	0.267	0.302
	2	0.407	0.021	0.386	0.444	0.350	0.019	0.326	0.376	0.297	0.020	0.262	0.321	0.182	0.035	0.147	0.217
	3	0.370	0.016	0.350	0.413	0.323	0.010	0.306	0.336	0.285	0.019	0.254	0.301	0.171	0.010	0.161	0.181
	4	0.403	0.015	0.383	0.420	0.361	0.010	0.348	0.380	0.325	0.013	0.301	0.340	0.273	0.019	0.242	0.289
	5	0.341	0.003	0.337	0.347	0.332	0.002	0.328	0.336	0.323	0.003	0.318	0.327	0.303	0.010	0.291	0.309
	6	0.408	0.015	0.389	0.438	0.358	0.010	0.343	0.377	0.310	0.020	0.282	0.327	0.243	0.014	0.234	0.264
	7	0.515	0.044	0.473	0.580	0.404	0.020	0.373	0.441	0.335	0.018	0.304	0.361	0.252	0.030	0.206	0.287
30–50 cm	1	0.421	0.011	0.404	0.439	0.384	0.007	0.376	0.400	0.366	0.006	0.350	0.373	0.316	0.014	0.300	0.328
	2	0.427	0.014	0.409	0.453	0.386	0.010	0.374	0.404	0.359	0.009	0.344	0.371	0.318	0.031	0.272	0.337
	3	0.347	0.011	0.336	0.363	0.324	0.008	0.311	0.333	0.297	0.011	0.283	0.306	0.219	0.010	0.209	0.229
	4	0.391	0.012	0.376	0.401	0.360	0.007	0.350	0.371	0.340	0.007	0.326	0.347	0.309	0.013	0.285	0.320
	5	0.374	0.004	0.368	0.382	0.356	0.003	0.353	0.361	0.347	0.003	0.341	0.350	0.333	0.005	0.327	0.338
	6	0.431	0.008	0.421	0.446	0.408	0.007	0.398	0.418	0.377	0.009	0.359	0.388	0.337	0.017	0.320	0.354
	7	0.432	0.008	0.425	0.444	0.415	0.006	0.404	0.421	0.387	0.006	0.376	0.398	0.349	0.009	0.340	0.358
0–50 cm	1	0.433	0.017	0.413	0.469	0.388	0.009	0.370	0.401	0.346	0.009	0.331	0.359	0.311	0.012	0.294	0.327
	2	0.411	0.015	0.383	0.434	0.355	0.013	0.339	0.371	0.314	0.012	0.293	0.332	0.221	0.010	0.211	0.231
	3	0.384	0.008	0.374	0.398	0.342	0.013	0.319	0.361	0.289	0.007	0.280	0.297	0.210	0.010	0.200	0.220
	4	0.423	0.006	0.416	0.430	0.367	0.015	0.344	0.392	0.318	0.010	0.305	0.337	0.273	0.011	0.261	0.282
	5	0.412	0.010	0.400	0.433	0.384	0.007	0.375	0.394	0.361	0.006	0.353	0.371	0.325	0.016	0.309	0.341
	6	0.432	0.012	0.417	0.453	0.392	0.010	0.374	0.410	0.349	0.010	0.333	0.363	0.281	0.023	0.247	0.297
	7	0.437	0.018	0.416	0.464	0.387	0.012	0.368	0.403	0.344	0.010	0.330	0.360	0.296	0.015	0.279	0.313

[a] Troxler Neutron Surface and Depth Probes (Hollinger and Isard, 1994).

(shape parameter that account for the pore size distribution), θ_{res} (residual soil moisture content), θ_{sat} (saturated soil moisture content), and K_{sat} (saturated hydraulic conductivity) indicates the Mualem-van Genuchten parameters. The soil moisture dynamics were estimated by the SWAP model (van Dam et al., 1997) using the soil hydraulic parameters derived by GA. Then, we tested the transposability of data assimilation scheme and HMMGA for calibration and validation under different weather conditions from 1994 to 1996 at the Bondville site in Illinois.

In the proposed HMMGA process, we generated multiple realizations (r) using a re-sampling technique (Efron, 1982) for uncertainty analysis of weekly soil moisture dynamics. Also, the multi-populations (from initial random seed numbers of −1000, −950, and −750) were used for assessing uncertainties of the near-surface soil moisture assimilation scheme. The determination correlation (R^2) and root mean square error (*RMSE,* $cm^3\ cm^{-3}$) were used for evaluating uncertainties,

$$R^2 = \frac{\left(\frac{1}{W}\sum_{w=1}^{W}(X_w-\overline{X})(Y_w-\overline{Y})\right)^2}{\frac{1}{W}\sum_{w=1}^{W}(X_w-\overline{X})^2\frac{1}{W}\sum_{w=1}^{W}(Y_w-\overline{Y})^2} \quad (14)$$

$$RMSE = \sqrt{\frac{\sum_{w=1}^{W}(X_w-Y_w)^2}{W}} \quad (15)$$

where R^2 is the determination correlation, X is the observed dataset, and Y is the simulated dataset.

3. Results and discussions

Fig. 3 shows the observed rainfall occurrence probabilities across the multiple weather stations with the Digital Elevation Models (DEMs) for the Oklahoma (130 km × 130 km) and Illinois (300 km × 500 km) domains. The rainfall occurrence probabilities (0.31–0.36) in Illinois were usually higher than those (0.26–0.29) in Oklahoma. Overall, the rainfall occurrences showed relatively small variations across the sites 1–7 in both domains. The topographies of the two domains were relatively flat and thus, the rainfall occurrences were not influenced as much by topography of the domains. The observed rainfall occurrence probabilities at the weather stations in the Oklahoma (relatively smaller) domain were correlated among the weather stations (with R^2 > 0.378–0.709), but the Illinois domain had relatively lower correlations (R^2: 0.050–0.350) between sites as shown in Fig. 4. It showed that spatial dependence of rainfall occurrences were significantly affected by the distance between the stations. In other words, the capability of HMM in simulating the multivariate time series of rainfall occurrences is influenced by DEMs, how closely the weather stations are located across the spatial domain, etc. Overall, the statistics (average and standard deviation) of soil moisture measurements with sandy loam (site 2-SPEN and site 4-ACME) and loam (site 5-WASH) in the Oklahoma region were relatively lower than those of other sites, while the statistics of soil moisture measurements, which have similar soil texture (silt loam), in the Illinois region showed relatively less variations across the sites 1–7.

The plots in Fig. 5 represent the performance (maximum log-likelihood) of HMMGA for the Oklahoma and Illinois domains. The stepwise improvements in searched solutions show that the log-likelihood maximum fitness derived by the HMMGA was instantly converged during the initial generation (max: 30). This result demonstrated that HMM can train the parameter set (Θ from Eq. (6)) well through the optimized initial parameters by the GA.

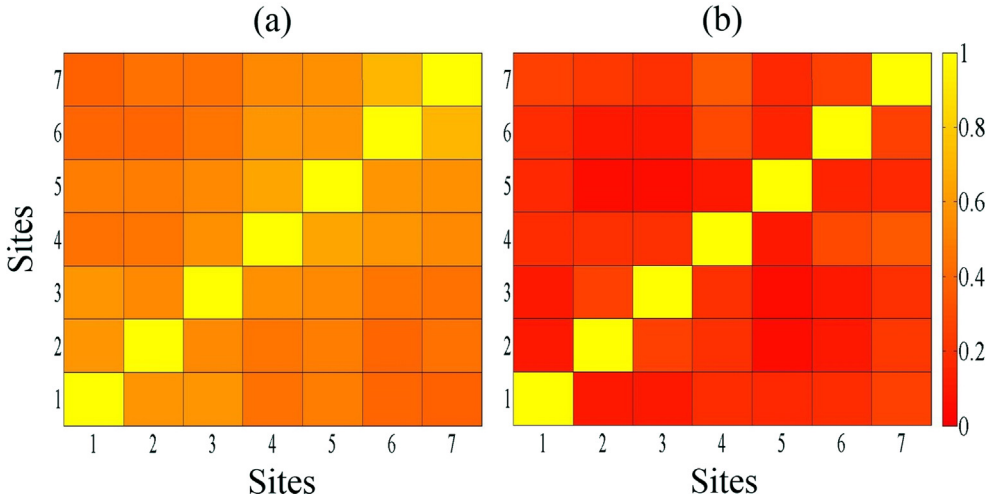

Fig. 4. Determination coefficient (R2) of observed rainfall occurrence probabilities between the multiple locations (sites 1–7) under different hydro-climatic regions; (a) Oklahoma domain and (b) Illinois domain.

3.1. Oklahoma domain

The (hidden) weather state transition probabilities for the Oklahoma domain (130 km × 130 km) are given in Table 4. The rows (*i*) represent the present weather states (state 1: wet, state 2: relatively wet, state 3: relatively dry, and state 4: dry conditions) at time *t* whereas the columns (*j*) represent the expected weather states at time $t + 1$.

Given a wet condition at time *t*, the probability of observing a wet condition at time $t + 1$ is highest (at 0.455) while the probability of transitioning to a dry state is 0.318. A relatively wet condition at time *t* also has the relatively high probabilities of observing wet (0.346) and dry conditions (0.375) at time $t + 1$ whereas relatively dry and dry conditions at time *t* has the high transition probabilities with a dry condition (0.679 and 0.753) at time $t + 1$. Overall, the weather state conditions at time *t* had the relatively higher transition probabilities with wet and dry conditions at time $t + 1$ than those of the relatively wet and dry conditions. It might indicate that moisture on the soil surface at these sites quickly run dry since the short-term rainfall durations. Further, the direct transitions between states 2 and 3 were rare, which meant that states 2 and 3 play the role of intermediaries.

Fig. 6 shows the comparison of observed rainfall occurrence probabilities (average over the study sites) and optimal state sequences (weather states) derived by the HMMGA during the simulation period (90 days from March 01 for each year during 1995–2009). Overall, the optimal state sequences had good match (R^2: 0.894) with the observed rainfall occurrence probabilities along the time series. This result showed that the derived sequences were capable of representing the weather state conditions. In an identical way, we tested the model performances of HMM only and HMMGA using three different random number generators (seeds: 1000, 2000, and 3000) in Oklahoma. As the HMMGA-based model outputs (R^2: 0.611–0.894) showed the higher correlations than those (R^2: 0.154–0.265) of HMM only (note that the HMM-based results were not shown in the text), we confirmed that our approach could provide more reliable results.

Fig. 7a–g shows the comparison of yearly observed and simulated rainfall occurrence probabilities for the individual sites. The yearly simulated probabilities matched with the observed datasets (R^2:

Table 4
Predicted transition probabilities for weather states under different hydro-climatic regions.

Study sites	State at time *t*	State at time $t + 1$			
		State 1	State 2	State 3	State 4
Oklahoma domain	State 1	0.455	0.140	0.087	0.318
	State 2	0.346	0.094	0.185	0.375
	State 3	0.167	0.138	0.016	0.679
	State 4	0.134	0.077	0.036	0.753
Illinois domain	State 1	0.377	0.131	0.341	0.151
	State 2	0.246	0.294	0.067	0.393
	State 3	0.191	0.091	0.358	0.360
	State 4	0.096	0.090	0.165	0.649

State 1: wet condition; State 2: relatively wet condition; State 3: relatively dry condition; State 4: dry condition.

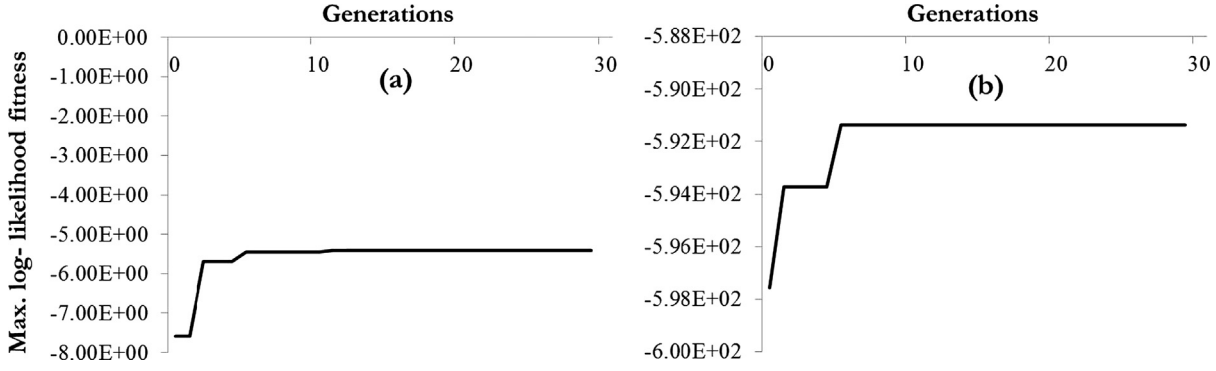

Fig. 5. Fitness of hidden Markov model based with a genetic algorithms (HMMGA) under different hydro-climatic regions; (a) Oklahoma domain and (b) Illinois domain. Note that a log-likelihood value closer to zero denotes better performance of the HMMGA.

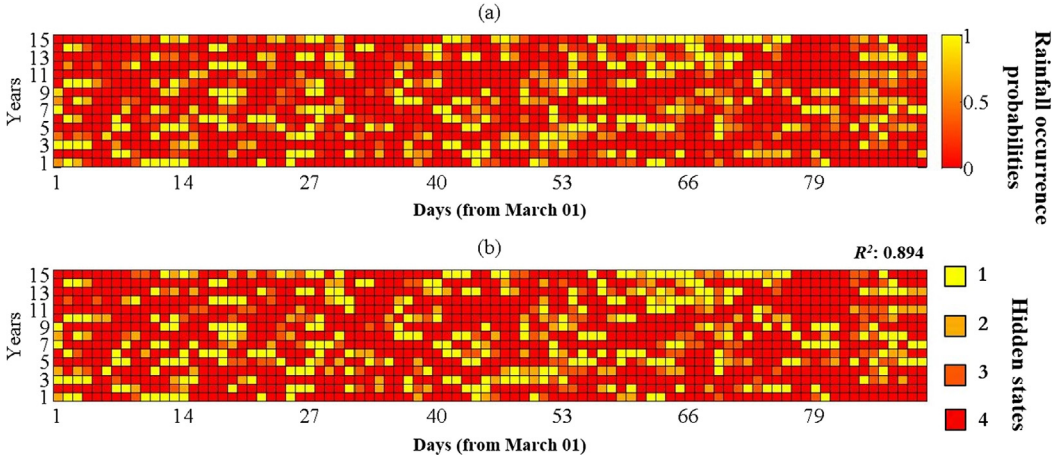

Fig. 6. (a) The observed rainfall occurrence probabilities and (b) optimal state sequences derived by the HMMGA (representing weather wetness conditions comprised of hidden states 1–4) during 1995–2009 (years 1–15) in Oklahoma. State 1: wet condition; State 2: relatively wet condition; State 3: relatively dry condition; State 4: dry condition.

0.119–0.457 and *RMSE*: 0.074–0.084). However, when we compared the observed and predicted rainfall occurrence probabilities (averaged for all years) for the individual weather stations at the entire spatial domain (Fig. 7h), the results (R^2: 0.615 and *RMSE*: 0.012) were highly correlated. This implies that our proposed scheme could predict long term rainfall occurrences well at both the spatial and temporal scales, especially better for the spatial domain.

Similarly, we predicted the time averaged (2003–2009) soil moisture (0–5 cm) for the individual sites, as shown in Fig. 8. The predicted soil moisture matched quite well with the measurements (R^2: 0.997 and *RMSE*: 0.005). The observed and estimated soil moistures for the sites 2, 4, 5, and 7 were relatively lower than those of sites 1, 3, and 6. When we compared the (observed/simulated) rainfall occurrence probabilities with the soil moisture predictions, the soil moisture estimates did not correspond to the rainfall patterns. This supported that the soil moisture was affected by not only the weather conditions, but also by the land surface conditions (i.e., soil textures, vegetation covers, topography, etc.).

3.2. Illinois domain

We tested the HMMGA approach for the Illinois domain (300 km × 500 km). The (hidden) state transition probabilities given in Table 4 show slightly different behavior compared to the results of the

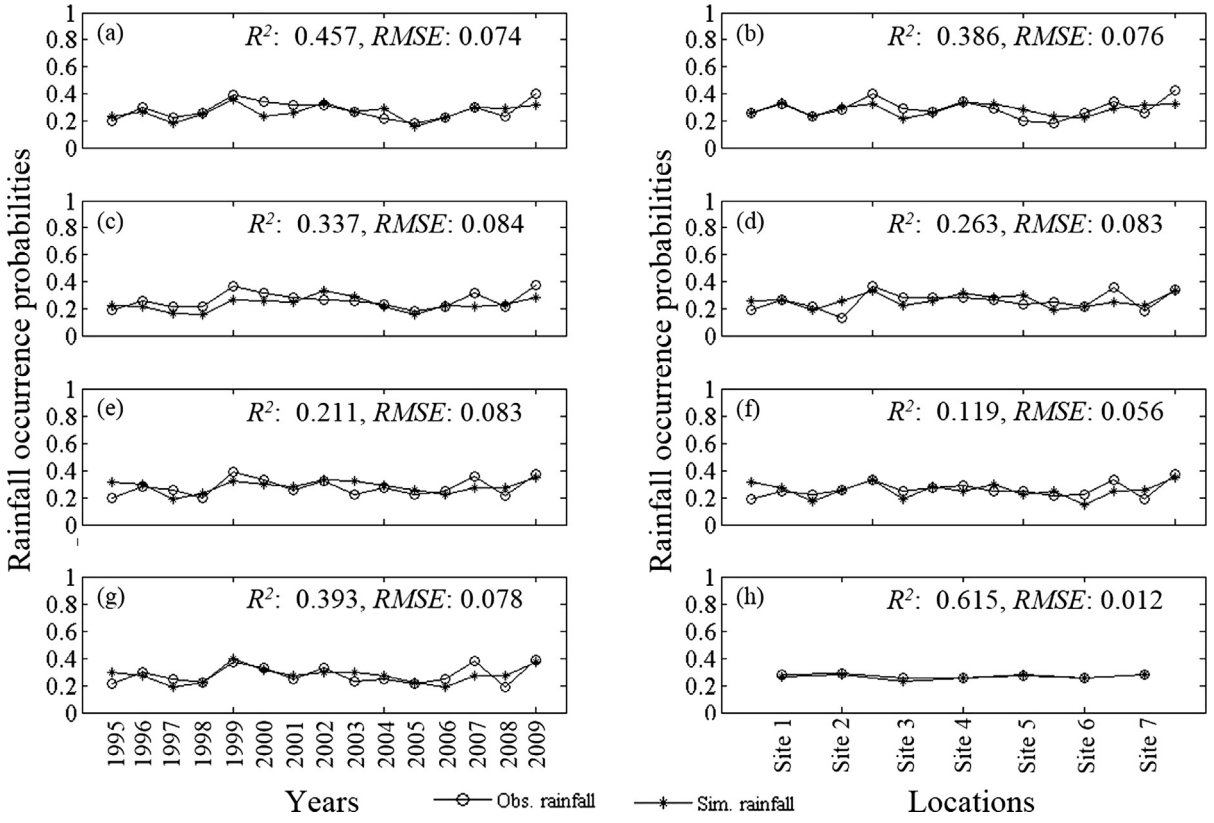

Fig. 7. (a–g) Yearly observed and simulated rainfall occurrence probabilities (1995–2009) for each weather station (sites 1–7) at the temporal domain and (h) average probabilities of observed and simulated rainfall occurrences for all years at the spatial domain in the Oklahoma domain.

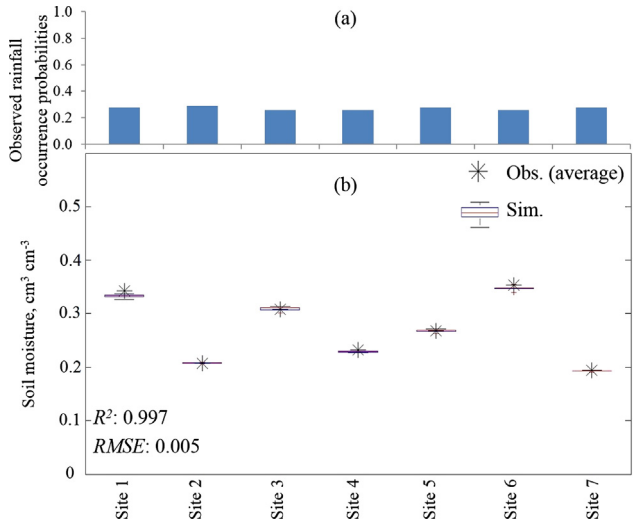

Fig. 8. (a) Observed rainfall occurrence probabilities (averaged for all years) and (b) measured and predicted soil moisture values* (averaged for all years, $cm^3\ cm^{-3}$) at the soil depth of 0–5 cm in the network of multiple weather locations for the Oklahoma domain during the simulation period (2003–2009). *On each box indicating the simulated soil moisture, the central line is the median value of data, the edges of the box are the 25th and 75th percentiles, and the whiskers extend to the most extreme data points.

Oklahoma domain.

A wet condition at time t had relatively high probabilities of observing wet (0.377) and relatively dry (0.341) conditions at time $t + 1$. A relatively wet condition at time t was evenly transitioned to other conditions (0.246–0.393) at time $t + 1$, except of a relatively dry condition (0.067). Given a relatively dry condition at time t, the probabilities of observing relatively dry (0.358) and dry (0.360) conditions at time $t + 1$ were higher than others, but a dry condition at time t still had the highest self-transition probability (0.649) at time $t + 1$. Overall, the probabilities of observing a dry condition at time $t + 1$ in Illinois were lower than those of Oklahoma indicating that the Illinois sites had more rainfall events.

Fig. 9 exhibits the comparison of observed rainfall occurrence probabilities and estimated optimal state sequences. The estimated optimal state sequences matched quite well (R^2: 0.794) with the observed rainfall occurrence probabilities, similar to the results of the Oklahoma domain (shown in Fig. 6). This result indicated that the HMMGA performs well in predicting the optimal state sequences (weather states) for the Illinois (relatively larger) domain also, although more uncertainties exist.

In Fig. 10a–g, the yearly predicted rainfall occurrence probabilities (R^2: 0.006–0.215 and *RMSE*: 0.063–0.084) for the sites 1–7 derived by the HMMGA were less identifiable with the observations than those (R^2: 0. 119–0.457 and *RMSE*: 0.074–0.084) of the Oklahoma (relatively smaller) domain in Fig. 7a–g, especially for the sites 5 and 7 (R^2: 0.010 and 0.006, *RMSE*: 0.084 and 0.077). It may indicate that the predictions were considerably influenced by the distance between the weather stations. The time averaged probabilities (R^2: 0.925 and *RMSE*: 0.006) for all years in the individual sites were fairly matched with the observations (Fig. 10h) at the spatial domain, although the predicted rainfall probabilities were slightly underestimated. In line with the results for the Oklahoma domain, it was found that the stochastic approach may be more robust in predicting rainfall occurrences across the entire spatial domain rather than predicting daily time series of rainfall occurrences at individual weather stations. The statistics of rainfall predictions have some uncertainties in modeling for the Illinois domain, but overall the predicted results tend to be similar with the observations.

Fig. 11 shows the weekly (7 days) measured and predicted soil moisture estimates ($cm^3\ cm^{-3}$) in the soil depth of 0–50 cm derived by our HMMGA approach and also from the near-surface soil moisture data assimilation scheme (Ines and Mohanty, 2008a) at Site 1 (Bondville) in Illinois for calibration (1994) and validation (1996). The model outputs for HMMGA (R^2: 0.699 and *RMSE*: 0.049) and physically-based assimilation scheme (R^2: 0.449 and *RMSE*: 0.028) matched the measurements for calibration in 1994, although the HMMGA-based results were slightly underestimated. We tested the transposability of our HMMGA approach and data assimilation scheme adapting the physically-based SWAP model (Ines and Mohanty, 2008a) across time for validation (1996). The results of physically-based data assimilation scheme (R^2: 0.023 and *RMSE*: 0.049) for validation (1996) had relatively more variations compared to those (R^2: 0.688 and *RMSE*: 0.055) of HMMGA. It demonstrated that physically-based hydrological models integrated with optimization schemes have limitations in reproducing soil moisture dynamics in the context of transposability with different weather conditions because the data assimilation scheme might be inclined to optimize the model input parameters under the specific weather and environmental conditions. The HMMGA-based model outputs were also slightly underestimated compared to the measurements for both the calibration and validation periods, but the results were consistent through the periods. Overall, our approach had uncertainties/limitations in predicting highest/lowest state probabilities corresponding to extreme rainfall events along time. These

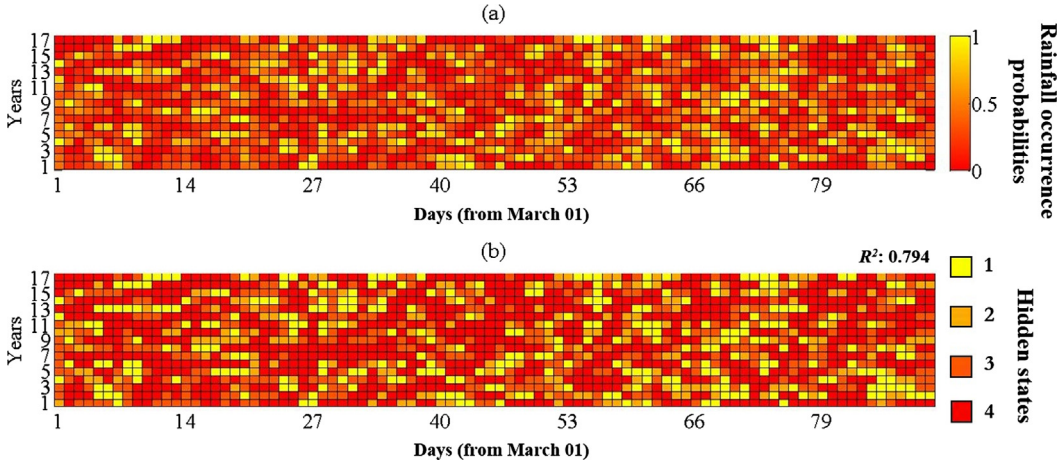

Fig. 9. (a) The observed rainfall occurrence probabilities and (b) optimal state sequences derived by the HMMGA (representing weather wetness conditions comprised of hidden states 1–4) during 1994–2010 (years 1–17) over the multiple weather locations in Illinois. State 1: wet condition; State 2: relatively wet condition; State 3: relatively dry condition; State 4: dry condition.

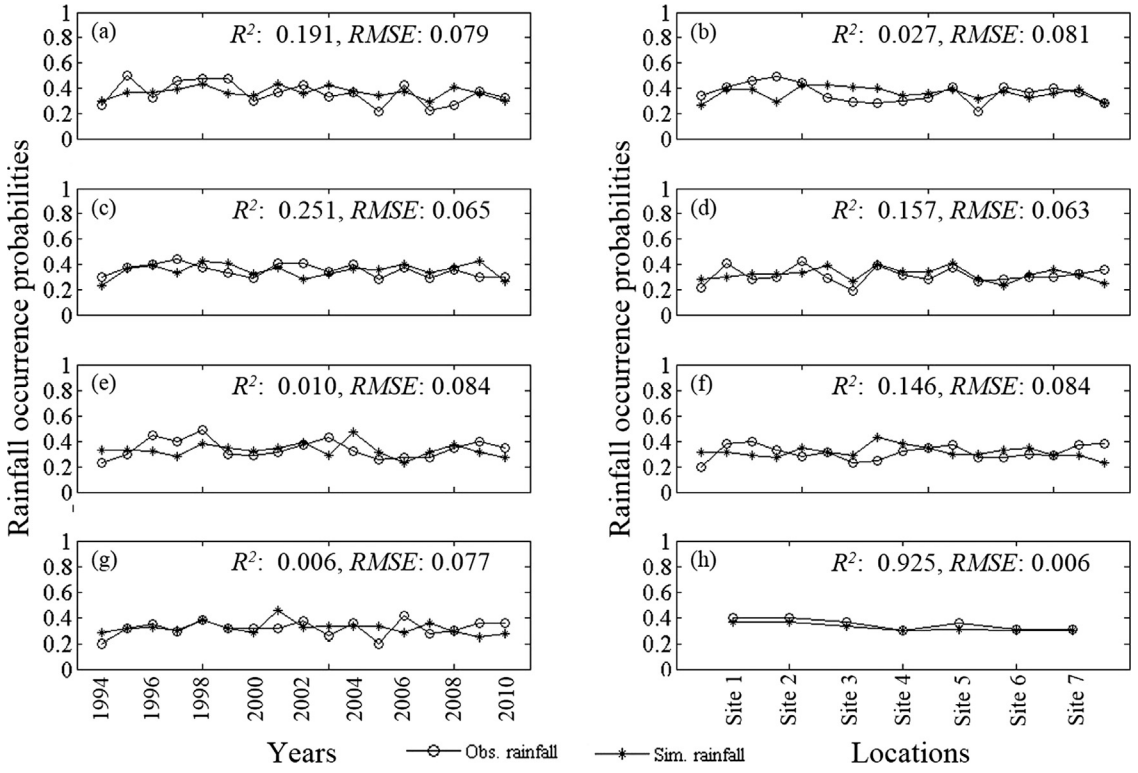

Fig. 10. (a–g) Yearly observed and simulated rainfall occurrence probabilities (1994–2010) for each weather station (sites 1–7) at the temporal domain and (h) average probabilities of observed and simulated rainfall occurrences for all years at the spatial domain in the Illinois domain.

uncertainties might be common issues in stochastic approaches that need to be improved in future. At least, these results demonstrated that our approach can estimate the reliable weekly model outputs with predictable uncertainties and provide advantages for the good transposability under different weather conditions.

We predicted the time averaged (1994–2000) soil moisture at the near-surface (0–10 cm) and in the sub-surface layers (10–30, 30–50, and 0–50 cm) at each of the weather stations (Fig. 12). The predicted soil moistures (R^2: 0.456–0.842 and *RMSE*: 0.024–0.041) at the various soil depths matched well with the measurements. The near-surface (0–10 cm) soil moisture estimates were more variable than those of the sub-surface layers. Mostly, the predicted soil moisture values have uncertainties compared to the measurements, especially for the sites 2 and 3. These uncertainties may be caused by the underestimations of predicted rainfall occurrence probabilities (Fig. 10h). Fig. 13 shows the calibration (1994–2000) and validation (2001–2004) results for soil moisture (at the soil depth of 0–50 cm). Overall, the predicted soil moisture estimates (R^2: 0.240 and *RMSE*: 0.051) during the validation (2001–2004) period matched the observations with uncertainties compared to those (R^2: 0.770 and *RMSE*: 0.040) for the calibration (1994–2000), while the soil moisture predictions of sites 2 and 3 for the validation have more bias, especially for site 2 causing the relatively low correlation. These findings are quite important in this study, because our proposed methodology can provide an attractive alternative in forecasting long-term soil moisture for the future using only climate change scenarios (i.e., global climate models-GCMs, regional climate models-RCMs, etc.) and historical soil moisture measurements (available). In this regard, this approach could be quite useful for developing/improving efficient agricultural/water resources management in the future.

Generally, the soil moisture predictions for the Illinois domain had more uncertainties than those of the Oklahoma (relatively smaller)

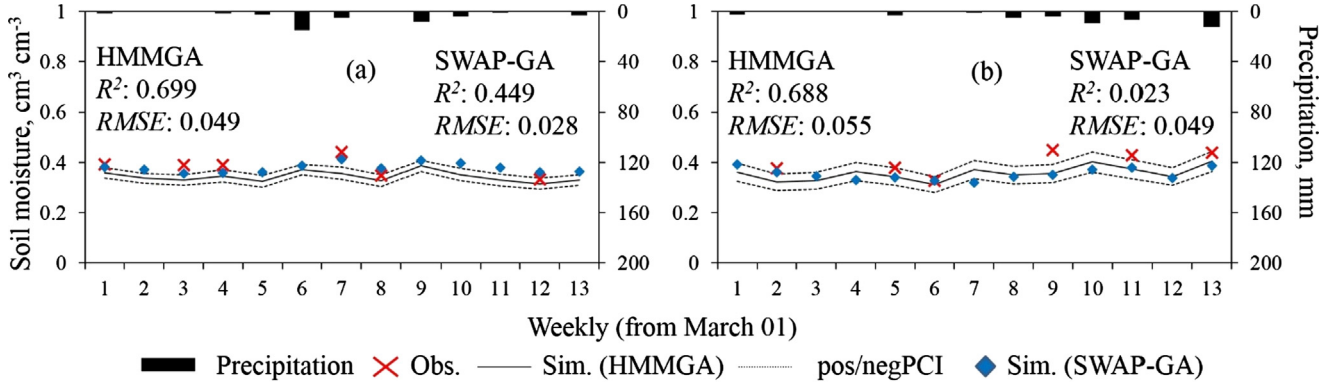

Fig. 11. Test of the transposability of the HMMGA[a] and near-surface data assimilation scheme[b] in time for predicting the soil moisture dynamics in the soil depth of 0–50 cm at the Bondville site (Illinois); (a) Calibration in 1994 and (b) Validation in 1996. [a]Realizations = 30; Random number generator seeds (3000); Max. generation = 30; Population = 10 [b]Random number generator seeds (−1000, −950, and −750); Max generation = 500; Population = 10.

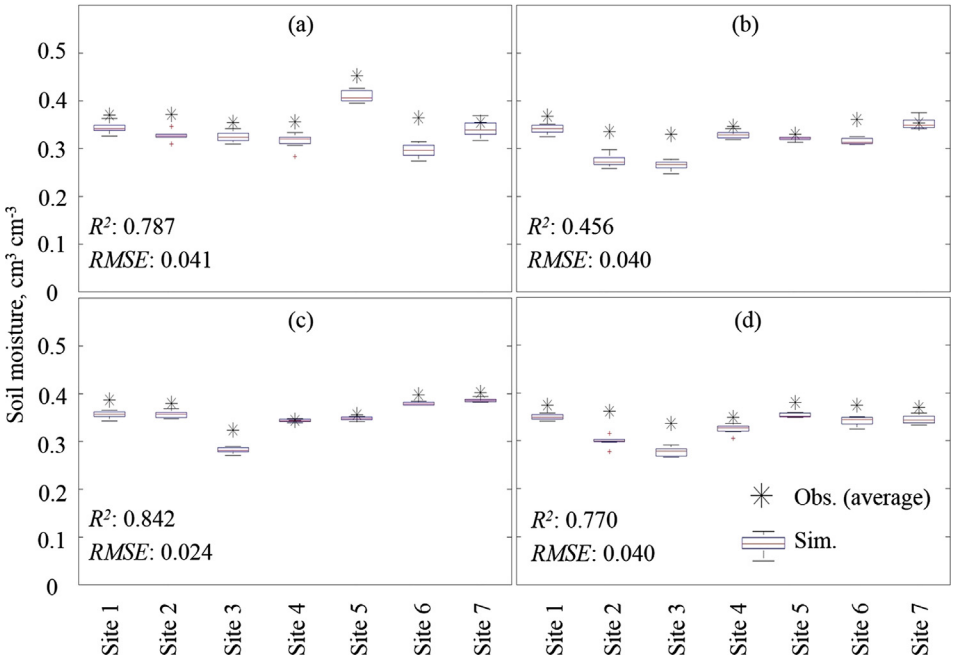

Fig. 12. Measured and predicted soil moisture dynamics* (averaged for all years, $cm^3\ cm^{-3}$) at various soil depths in the network of multiple weather locations for the Illinois domain during the simulation period (2003–2009); (a) 0–10 cm, (b) 10–30 cm, (c) 30–50 cm, and (d) 0–50 cm. *On each box indicating the simulated soil moisture, the central line is the median value of data, the edges of the box are the 25th and 75th percentiles, and the whiskers extend to the most extreme data points.

domain. Transition probabilities estimated from rainfall occurrences across multiple locations might be influenced by spatial extents. Thus, it is usual that HMMGA-based soil moisture predictions had more uncertainties at larger extents. Although the results for the Illinois domain had more bias compared to the measurements, the HMMGA approach still predicted the soil moisture well at the temporal (weekly) and spatial scales with uncertainties. The data assimilation scheme using the physical-based hydrological (SWAP) model also has uncertainties (i.e., due to the model structure, parameter uncertainties, measurement and user's modeling skill errors, transposability, etc.). When we considered these limitations mentioned above and complexity of applications, our suggested approach can provide an easy application and minimize uncertainties for soil moisture predictions using only precipitation and historical soil moisture measurements existing in the real world conditions.

4. Conclusion

In this study, we developed a genetic algorithm-based hidden Markov model (HMMGA) for predicting spatially and temporally distributed soil moisture using only precipitation and historical soil moisture measurements across multiple weather stations and tested the transposability of HMMGA in time during calibration and validation under different weather conditions. Rainfall occurrence probabilities reproduced by HMMGA were also analyzed. HMMGA was used to estimate most likely (optimal) hidden (weather) state sequences (representing weather states comprised of the four hidden states-wet, relatively wet, relatively dry, and dry conditions) based on precipitation history. Historical soil moisture data measured at different soil depths (0–5, 0–10, 10–30, 30–50, and 0–50 cm) were categorized based on the weather state conditions. Then, statistics (mean and standard deviation) of categorized soil moisture data were determined. Using the HMMGA-based weather states and historical categorized soil moisture statistics, soil moisture dynamics were estimated. Two different domains in Oklahoma (1995–2009, 130 km × 130 km) and Illinois (1994–2010, 300 km × 500 km) under different hydro-climatic locations were selected for testing this approach.

HMMGA estimated well the optimal hidden state sequences for the Oklahoma (R^2: 0.894) and Illinois (R^2: 0.794) domains compared to the observed rainfall occurrence probabilities. The yearly predicted rainfall occurrence probabilities at the individual weather stations for the Oklahoma domain had a better match (R^2: 0.119–0.457 and *RMSE*: 0.056–0.084) with observations, but the Illinois domain had more variations (R^2: 0.006–0.251 and *RMSE*: 0.063–0.084) indicating that

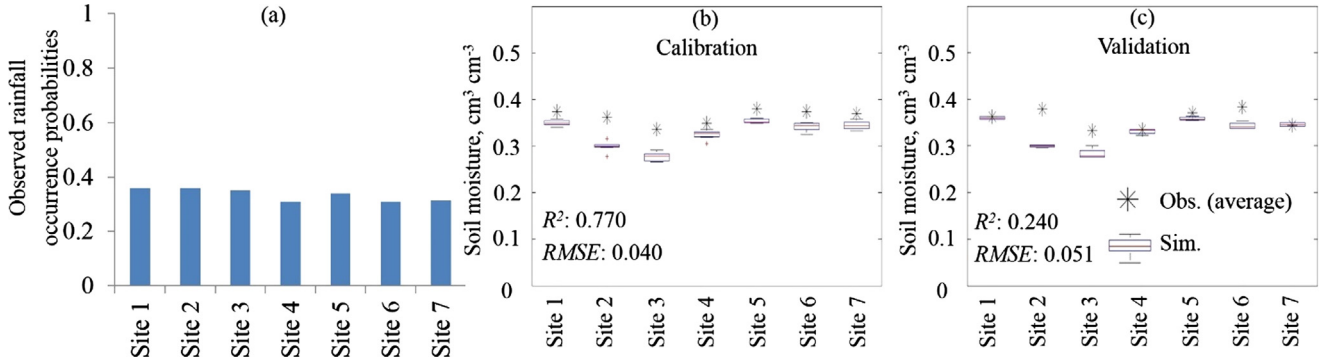

Fig. 13. Measured and predicted soil moisture dynamics* (averaged for all years, $cm^3\ cm^{-3}$) at the soil depth (0–50 cm) in the network of multiple weather locations for the Illinois domain during the calibration and validation periods; (a) observed rainfall occurrence probabilities, (b) calibration (1994–2000), and (c) validation (2001–2004). *On each box indicating the simulated soil moisture, the central line is the median value of data, the edges of the box are the 25th and 75th percentiles, and the whiskers extend to the most extreme data points.

the HMMGA processes are dependent on the distance between the weather stations. The average rainfall occurrence predictions for all years across the individual sites matched the observations well (R^2: 0.615 and *RMSE*: 0.012 for the Oklahoma domain and R^2: 0.925 and *RMSE*: 0.006 for the Illinois domain). This result showed that HMMGA provided better rainfall occurrence predictions across a spatial domain than daily rainfall occurrence predictions in time. The averaged soil moisture dynamics for all seven years at the Oklahoma domain matched almost perfectly (R^2: 0.997 and *RMSE*: 0.005) with the measurements. The weekly (7 days) averaged model outputs (R^2: 0.699 and *RMSE*: 0.049) for HMMGA were more identifiable with the measured data in Illinois, although the HMMGA-based results were slightly underestimated compared to those (R^2: 0.449 and *RMSE*: 0.028) of the data assimilation scheme. However, the data assimilation-based validation (R^2: 0.688/0.023 and *RMSE*: 0.055/0.049 for HMMGA/data assimilation) results in 1996 had more uncertainties compared to those of calibration in 1994. Although uncertainties exist, these results demonstrated that the stochastic approach could provide relatively reliable model outputs at the weekly time intervals and relatively good transposability under different weather conditions. Also, the average soil moisture dynamics at the different soil depths (0–10, 10–30, 30–50, and 0–50 cm) performed fairly well (R^2: 0.456–0.842 and *RMSE*: 0.024–0.041) compared to the measurements, although uncertainties were evident in the predictions. Generally, the soil moisture predictions had slightly different trends compared to the rainfall occurrence probabilities for both domains. These findings showed that the predicted soil moisture estimates were influenced by not only the precipitation, but also the land surface conditions (e.g., topography, soil textures, vegetation covers, etc.). Thus, this scheme which uses rainfall and historical soil moisture data performed relatively well in predicting soil moisture, even though our stochastic-based HMMGA has limitations in capturing temporal dependence of soil moisture affected by the land surface conditions. The results of calibration and validation supported the robustness of this approach in predicting the soil moisture dynamics at the multiple locations. Thus, our approach can be used to predict soil moisture dynamics in the future using only climate change scenarios and soil moisture history existing. Currently, our approach was only applied to the study areas, which have the soil moisture history, but HMMGA can be also extended to various regions through continental/global soil moisture database (i.e., International Soil Moisture Network-ISMN, etc.). Furthermore, pixel-scale soil moisture estimates can be obtained by combining satellite-scale precipitation (TRMM/GPM) and soil moisture (AMSR-E, AMSR2, SMOS, SMAP, etc) products with our approach, although uncertainties existing in RS data (due to the discrepancy of spatio-temporal scales, short-term RS precipitation and soil moisture data indicating less than 20 years scanned by satellites at the global scale, etc.) might be included in satellite-scale soil moisture estimates. As the predicted results at the spatial domain matched quite well with the observations, our approach can be also used for downscaling both remotely sensed precipitation (Robertson et al., 2004) and soil moisture (Kwon et al., 2017) products. Overall, our findings support that our approach provides easy convenience (avoiding intensive input parameters for physically-based hydrological models) in application to fields and minimizes not only user's skills in determining initial parameter values, but also model structural (i.e., parameterizations, initial/bottom boundary conditions, governing equation, transposability, etc.) and measurement (i.e., land covers, soil information, etc.) errors. Furthermore, when we consider that the availability of RS soil moisture footprints are limited across forest regions, our proposed scheme may provide an alternative in predicting soil moisture dynamics using TRMM/GPM with statistics of historical soil moisture (at forest regions). However, HMMGA still requires to be improved in predicting daily soil moisture. This limitation might be resolved with increasing the number of hidden states under the assumption that long-term daily soil moisture measurements are available in large areas. Thus, the newly developed algorithm could be useful for agricultural/water resources management in an efficient and economical manner.

Acknowledgements

The research was funded by NASA-THP (NNX08AF55G, NNX09AK73G, NNX16AQ58G) and NSF (CMG/DMS) grants (0621113, 0934837). Daily precipitation datasets were obtained from the Oklahoma mesonet stations (1995–2009) in Oklahoma (https://www.mesonet.org/) and the International Soil Moisture Network (1994-2010) in Illinois (http://ismn.geo.tuwien.ac.at/networks/icn/). Daily volumetric soil moisture were collected from the mesonet stations (Oklahoma). At the Illinois locations, biweekly volumetric soil moisture data at different soil depths were collected from by the Illinois State Water Survey (https://www.isws.illinois.edu/).

References

Albergel, C., Rüdiger, C., Carrer, D., Calvet, J.-C., Fritz, N., Naeimi, V., Bartalis, Z., Hasenauer, S., 2008. An evaluation of ASCAT surface soil moisture products with in-situ observations in southwestern France. Hydrol. Earth Syst. Sci. Discuss. 5, 2221–2250. https://doi.org/10.5194/hessd-5-2221-2008.

Barrett, B., Petropoulos, G., 2013. Satellite remote sensing of surface soil moisture. Remote Sens. Energy Fluxes Soil Moisture Content 85–120. https://doi.org/10.1201/b15610-6.

Bouttier, F., Mahfouf, J.-F., Noilhan, J., 1993a. Sequential assimilation of soil moisture from atmospheric low-level parameters. Part I: Sensitivity and calibration studies. J. Appl. Meteoro. 32, 1335–1351.

Bouttier, F., Mahfouf, J.-F., Noilhan, J., 1993b. Sequential assimilation of soil moisture from atmospheric low-level parameters. Part II: Implementation in a mesoscale model. J. Appl. Meteoro. 32, 1352–1364.

Brocca, L., Hasenauer, S., Lacava, T., Melone, F., Moramarco, T., Wagner, W., Dorigo, W., Matgen, P., Martínez-Fernández, J., Llorens, P., Latron, J., Martin, C., Bittelli, M., 2011. Soil moisture estimation through ASCAT and AMSR-E sensors: an intercomparison and validation study across Europe. Remote Sens. Environ. 115, 3390–3408. https://doi.org/10.1016/j.rse.2011.08.003.

Carroll, D.L., 1996. Genetic algorithms and optimizing chemical oxygen-Iodine lasers. In: In: Wilson, H.B. (Ed.), Developments in Theoretical and Applied Mechanics, vol. 18. Univ. of Ala, Tuscaloosa, AL, pp. 411–424.

Casella, G., Berger, R.L., 2001. Statistical Inference, 2nd ed. Duxbury. ISBN 0-534-24312-6.

Crow, W.T., Berg, A.A., Cosh, M.H., Loew, A., Mohanty, B.P., Panciera, R., de Rosnay, P., Ryu, D., Walker, J.P., 2012. Upscaling sparse ground-based soil moisture observations for the validation of coarse-resolution satellite soil moisture products. Rev. Geophys. 50 (RG2002). https://doi.org/10.1029/2011RG000372.

Das, N.N., Entekhabi, D., Njoku, E.G., 2011. An algorithm for merging SMAP radiometer and radar data for high resolution soil moisture retrieval. IEEE Trans. Geosci. Remote Sens. 49 (5). https://doi.org/10.1109/TGRS.2010.2089526.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via EM algorithm. J. Roy. Stat. Soc., Series B (Methodological) 39 (1), 1–38.

Dumedah, G., Berg, A., Wineberg, M., 2012. Evaluating auto-selection methods used for choosing solutions from Pareto-optimal set: does non-dominance persist from calibration to validation phase? J. Hydro. Eng. 17 (1). https://doi.org/10.1061/(ASCE)HE.1943-5584.0000389.

Entekhabi, D., Rodriguez-Iturbe, I., 1994. Analytical framework for the characterization of the space-time variability of soil moisture. Adv. Water Resour. 17, 35–45.

Entekhabi, D., Asrar, G.R., Betts, A.K., Beven, K.J., Bras, R.L., Duffy, C.J., 1999. An agenda for land surface hydrology research and a call for the Second International Hydrological Decade. Bull. Am. Meteoro. Soc. 80 (10), 2043–2058.

Engman, T., 1991. Application of microwave remote sensing of soil moisture for water resources and agriculture. Remote Sens. Environ. 35, 213–226.

Efron, B., 1982. The Jackknife, the Bootstrap and Other Resampling Plans. Society for Industrial and Applied Mathematics, Philadelphia, PA.

Goldberg, D.E., 2002. The Design of Innovation: Lessons from and for Competent Genetic Algorithms. Kluwer Academic Publishers, Norwell, Mass.

Goldberg, D.E., 1989. Genetic Algorithms in Search and Optimization and Machine Learning. Addison-Wesley Publishing, New York.

Hughes, J.P., Guttorp, P., 1999. A non-homogeneous hidden Markov model for precipitation occurrence. J. Roy. Stat. Soc. 48C, 15–30.

Hollinger, S.E., Isard, S.A., 1994. A soil moisture climatology of Illinois. J. Climate 7, 822–833.

Hartigan, J., Wong, M., 1979. A k-means clustering algorithm. Appl. Stat. 28, 100–108.

Holtan, H.N., 1975. A concept for infiltration estimates in watershed engineering, USDA ARS Paper, pp. 41–51.

Holland, J.H., 1975. Adaption in Natural and Artificial System. University of Michigan press, Ann Arbor, MI.

Hou, A.Y., Kakar, R.K., Neeck, S., Azarbarzin, A.A., Kummerow, C.D., Kojima, M., Oki, R., Nakamura, K., Iguchi, T., 2014. The global precipitation measurement mission. Bull. Am. Meteorol. Soc. 95, 701–722. https://doi.org/10.1175/BAMS-D-13-00164.1.

Ines, A.V.M., Mohanty, B.P., 2009. Near-surface soil moisture assimilation for quantifying

effective soil hydraulic properties using genetic algorithm (II. Using airborne remote sensing during SGP97 and SMEX02). Water Resour. Res. 45, W01408. https://doi.org/10.1029/2008WR007022.

Ines, A.V.M., Mohanty, B.P., 2008a. Near-surface soil moisture assimilation for quantifying effective soil hydraulic properties using genetic algorithm: I. Conceptual modeling. Water Resour. Res. 44, W06422. https://doi.org/10.1029/2007WR005990.

Ines, A.V.M., Mohanty, B.P., 2008b. Near-surface soil moisture assimilation for quantifying effective soil hydraulic properties under different hydro-climatic conditions. Vadose Zone J. 7, 39–52.

Illston, B.G., Basara, J.B., Fisher, D.K., Elliott, R., Fiebrich, C.A., Crawford, K.C., Humes, K., Hunt, E., 2008. Mesoscale monitoring of soil moisture across a statewide network. Am. Meteoro. Soc. 25, 167–182. https://doi.org/10.1175/2007JTECHA993.1.

Ines, A.V.M., Honda, K., 2005. On quantifying agricultural and water management practices from low spatial resolution RS data using genetic algorithms: A numerical study for mixed-pixel environment. Adv. Water Resour. 28, 856–870.

Ines, A.V.M., Droogers, P., 2002. Inverse modeling in estimating soil hydraulic functions: a genetic algorithm approach. Hydrol. Earth Syst. Sci. 6 (1), 49–65.

Jain, A.K., Murty, M.N., Flynn, P.J., 1999. Data clustering: a review. ACM Comput. Surv. 31 (9), 264–323.

JAXA, 2013. GCOM-W1 SHIZUKU Data Users Handbook, first ed.

Kerr, Y.H., Waldteufel, P., Richaume, P., Wigneron, J.P., Ferrazzoli, P., Mahmoodi, A., Bitar, A.A., Cabot, F., Gruhier, C., Juglea, S.E., Leroux, D., Mialon, A., Delwart, S., 2012. The SMOS soil moisture retrieval algorithm. Geosci. Remote Sens. 50, 1384–1403. https://doi.org/10.1109/TGRS.2012.2184548.

Kirshner, S., 2005. Modeling of Multivariate Time Series using Hidden Markov Models. University of California, Irvine, CA Ph.D diss.

Krishnakumar, K., 1989. Microgenetic algorithms for stationary and nonstationary function optimization. SPIE Intelligent Control Adaptive Syst. 1196, 289–296.

Kummerow, C., Barnes, W., Kozu, T., Shiue, J., Simpson, J., 1998. The Tropical Rainfall Measuring Mission (TRMM) sensor package. J. Atmos. Ocean. Technol. 15, 809–817. https://doi.org/10.1175/1520-0426(1998) 015<0809:TTRMMT>2.0.CO;2.

Kwon, M., Kwon, H.H., Han, D., 2017;al., in press. A spatial downscaling of soil moisture from rainfall, temperature, and AMSR2 using a Gaussian-mixture nonstationary hidden Markov model. J. Hydrol. https://doi.org/10.1016/j.jhydrol.2017.12.015. in press.

Liang, X., Wood, E.F., Lettenmaier, D.P., 1996. Surface soil moisture parameterization of the VIC-2L model: Evaluation and modifications. Global Planet. Change 13, 195–206.

MacDonald, I.L., Zucchini, W., 1997. Hidden Markov and Other Models for Discrete-Valued Time Series. Chapman & Hall, Boca Raton, FL.

Mahfouf, J.F., 1990. A variational assimilation of soil moisture in meteorological models. In: International Symposium on Assimilation of Observations in Meteorology and Oceanography, Clermont Ferrand (France), pp. 249–254.

Mehrotra, R., Sharma, A., 2005. A nonparametric nonhomogeneous hidden Markov model for downscaling of multisite daily rainfall occurrences. J. Geophys. Res. 110 (D16108). https://doi.org/10.1029/2004JD005677.

Mohanty, B.P., Cosh, M., Lakshmi, V., Montzka, C., 2017. Soil moisture remote sensing: state-of-the-science. Vadose Zone J. 16 (1). https://doi.org/10.2136/vzj2016.10.0105.

Nazemi, A., Yao, X., Chan, A., 2006. Extracting a set of robust Pareto-optimal parameters for hydrologic models using NSGA-II and SCEM. IEEE Con. Evolu. Compu., IEEE, New York, pp. 1901–1908.

Oleson, K., Dai, Y., Bonan, G.B., Bosilovich, M., Dickinson, R., Dirmeyer, P., Hoffman, F., Houser, P., Levis, S., Niu, G.-Y., Thornton, P., Vertenstein, M., Yang, Z., Zeng, X., 2004. Technical Description of the Community Land Model (CLM), NCAR Technical Note, NCAR/TN-461 + STR, 10.5065/D6N877R0.

Pan, F., Peters-Lidard, C.D., Sale, M.J., 2003. An analytical method for predicting surface soil moisture from rainfall observations. Water Resour. Res. 39 (11). https://doi.org/10.1029/2003WR002142.

Peck, E.L., 1976. Catchment modeling and initial parameter estimation for the national weather service river forecast system. In: NOAA Technology Memorandum, NWS HYDRO-31, National Weather Service, Silver Spring, MD.

Peng, J., Loew, A., Merlin, O., Verhoest, N.E.C., 2017. A review of spatial downscaling of satellite remotely sensed soil moisture. Rev. Geophys. 55, 341–366. https://doi.org/10.1002/2016RG000543.

Robertson, A., Kirshner, S., Smyth, P., 2004. Downscaling of daily rainfall occurrence over northeast brazil using a Hidden Markov Model. J. Climate 17, 4407–4424.

Rabiner, L.R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77 (2), 257–286.

Seiller, G., Anctil, F., Perrin, C., 2012. Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions. Hydrol. Earth Syst. Sci. 16, 1171–1189. https://doi.org/10.5194/hess-16-1171-2012.

Shin, Y., Mohanty, B.P., 2013. Development of a deterministic downscaling algorithm for remote sensing soil moisture footprint using soil and vegetation classifications. Water Resour. Res. 49. https://doi.org/10.1002/wrcr.20495.

Shin, Y., Mohanty, B.P., Ines, A.V.M., 2012. Soil hydraulic properties in one-dimensional layered soil profile using layer-specific soil moisture assimilation scheme. Water Resour. Res. 48, W06529. https://doi.org/10.1029/2010WR009581.

Shin, Y., Mohanty, B.P., Ines, A.V.M., 2013. Estimating effective soil hydraulic properties using spatially distributed soil moisture and evapotranspiration. Vadose Zone J. 12 (3). https://doi.org/10.2136/vzj2012.0094.

van Dam, J.C., Huygen, J., Wesseling, J.G., Feddes, R.A., Kabat, P., van Waslum, P.E.V., Groenendjik, P., van Diepen, C.A., 1997. Theory of SWAP version 2.0: Simulation of water flow and plant growth in the soil-water-atmosphere-plant environment. Technical Document 45, DLO Winand Staring Centre, Wageningen, Netherlands.

Viterbi, A.J., 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Trans. Inf. Theory 13 (2), 260–269.